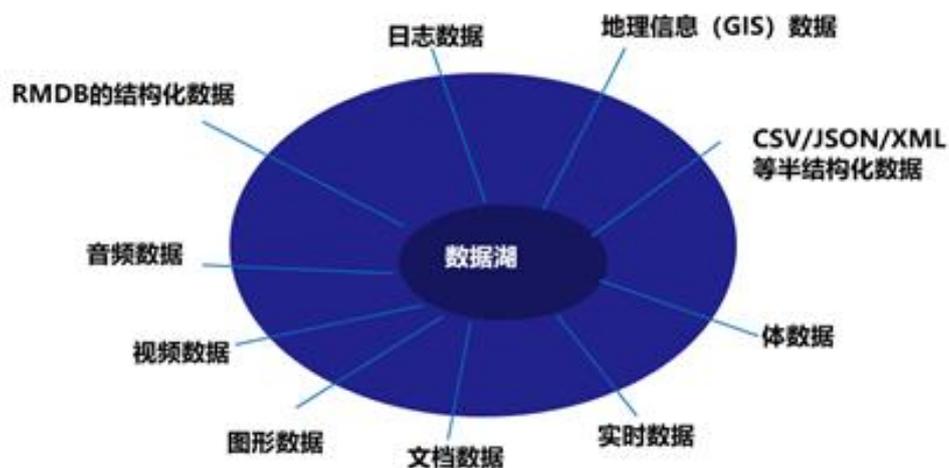


基于数据湖架构下的数据治理体系

来源	来自于百度文库
说明	本文主要介绍数据湖的概念，数据湖的定义及特点，分析了数据仓库与数据湖的异同并介绍了数据湖的架构体系；数据湖遇到的挑战，并给出了实践案例说明数据湖数据管理的 4 个能力，最后对数据湖未来进行了展望，希望对相关领域从业人员有所借鉴和帮助。



目录

基于数据湖架构下的数据治理体系.....	1
前言.....	3
一、 数据处理技术的发展趋势与挑战.....	3
1.1 数据管理面临的挑战和转变.....	4
1.2 数据湖的定义及发展需求.....	5
1.3 从数据库、数据仓库到数据湖演变趋势.....	6
1.4 数据仓库与数据湖差异.....	7
二、 数据湖的架构体系.....	9
2.1 数据湖架构体系.....	9
2.2 以 AWS 数据湖产品为例，实现数据管理 4 个能力.....	12
2.3 数据湖数据管理 4 个能力.....	15
三、 如何通过数据治理实现数据湖商业价值.....	16
3.1 数据湖遇到挑战.....	17
3.2 避免数据沼泽.....	17
3.3 数据智能化治理是数据湖实现价值必有之路.....	18
3.4 构建数据湖的数据治理体系相关思考.....	19
四、 Amazon Athena 和 AWS Glue 中国区域实践案例.....	21
4.1 ETL 服务为数据分析准备工作的自动化，大幅缩短数据准备时间.....	22
4.2 数据资源目录为数据湖提供智能化数据管理能力.....	22
4.3 交互式查询服务为数据湖提供高效、便捷服务能力.....	22
五、 数据湖的未来展望.....	23
六、 结束语.....	23

前言

随着大数据、人工智能、云计算、物联网等数字化技术的普及和广泛应用，传统的数据仓库模式，在快速发展的企业面前已然显得力不从心。数据湖，是可以容纳大量的原始数据的存储库和处理系统，已经成为企业应用大数据的重要工具。数据湖可以更好地支撑数据预测分析、跨领域分析、主动分析、实时分析以及多元化结构化数据分析，可以加速从数据到价值的过程，打造相应业务能力。而有效的数据治理才是数据资产形成的必要条件，同时数据治理是一个持续性过程，也是数据湖逐步实现数据价值的过程。未来在多方技术趋于融合，落地场景将不断创新，数据湖、数据治理或将成为新的技术热点。

本文第一章从数据管理面临的挑战与发展趋势分析了数据管理面临的三重挑战和三个转变以及数据管理技术的演进路线，引出了数据湖的概念；第二章给出了数据湖的定义及特点，分析了数据仓库与数据湖的异同并介绍了数据湖的架构体系；第三章分析了数据湖遇到的挑战，指出通过数据智能化治理是实现数据湖价值的必由之路，对构建数据湖治理体系进行了详细的分析；第四章给出了 Amazon Athena 和 AWS Glue 中国区域最佳实践案例，并以具体产品为例说明数据湖数据管理的 4 个能力，以帮助读者对数据湖管理技术有更为深入详细的认识；第五章对数据湖未来的发展进行了展望。希望对相关领域从业人员有所借鉴和帮助。

一、数据处理技术的发展趋势与挑战

在数字经济时代，应用程序在不断地产生并储存大量数据，而这些数据却无法及时被其他程序使用，导致“数据孤岛”产生。数据湖的诞生，不仅解决了“数据孤岛”的问题，还使企业获得更强的数据使用能力。作为存储企业原始数据的“大型仓库”，数据湖结合先进的数据科学与机器学习技术，不但能帮助企业构建更多优化后的运营模型，还能为企业提供预测分析、推荐模型等能力，促进企业增长。

1.1 数据管理面临的挑战和转变

随着大数据技术日益成熟，企业对经营管理风险防控、可视化监控、预测性分析和精细化管理提出了更高的要求，企业需要打破不同业务领域之间的壁垒，真正做到数据和业务流程的融会贯通，进一步挖掘数据价值，提升企业综合决策的能力，提高企业工作和管理效率。

(1) 数据管理面临的三个挑战

- 1) 数据仓库模式导致的烟囱式建设与数据需跨业务线广泛连接之间的挑战；
- 2) 传统数据库不能应对数据的增长，数据 ETL、数据建模工作的响应速度与数据反哺业务迭代创新之间的挑战；例如：移动互联网和物联网时代，产生了大量的网站数据，社交媒体数据，物联网设备数据等非结构化数据。导致数据仓库无法满足这些多元化的数据结构的存储和查询，以及非结构化和结构化数据的交叉分析。
- 3) 数据赋能与业务场景探索脱节的挑战。

(2) “数据+平台+应用”的新生态模式，实现数据分析三个方面的转变

1) 从统计分析向预测分析转变

从利用报表、图像展示等方式显示当前数据的内容概况，转变为利用人工智能、机器学习等手段预测数据的未来变化规律。

2) 从非实时向实时分析转变

经营决策者需更及时、快速的获取业务数据，以便及时根据市场变化调整经营策略。采用内存计算、消息队列等大数据分析方式实现实时分析。

3) 从结构化数据向多元化转变

利用自然语言处理、语音识别、图像识别等技术，将非结构化数据和结构化数据相结合，完善客户、供应商画像、设备的精准度，实现精准营销、物资供应和预防性维修等。

1.2 数据湖的定义及发展需求

数据湖(Data Lake)是 Pentaho 的 CTO James Dixon 提出来的，是一种数据存储理念——即在系统或存储库中以自然格式存储数据的方法。

目前，Hadoop 是最常用的部署数据湖的技术，所以很多人会觉得数据湖就是 Hadoop 集群。数据湖是一个概念，而 Hadoop 是用于实现这个概念的技术。数据湖到底是什么？业内并没有达成共识定义。我们先看看 Amazon AWS 把数据湖定义为：Amazon S3 存储、数据目录、数据冷备；并辅之以数据移动工具、数据分析工具、机器学习工具。注：为了维持定义的精确性，看英文原文如何描述。从 Amazon AWS 得到的解释：

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

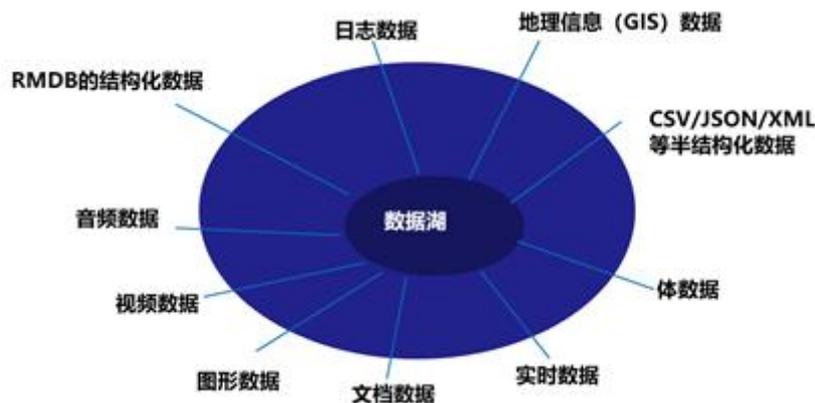


图 1.数据湖存储数据类型

数据湖是一个存储企业的各种各样原始数据的大型仓库，其中的数据可供存取、处理、分析及传输。数据湖从企业的多个数据源获取原始数据，并且针对不同的目的，同一份原始数据还可能有多多种满足特定内部模型格式的数据副本。



图 2.未经处理和包装的原生状态“水库”

(1) 数据湖是有一个中心化的存储，所有的数据以它本来的形式【包括结构化数据（关系数据库数据），半结构化数据（CSV、XML、JSON 等），非结构化数据（电子邮件，文档，PDF）和二进制数据（图像、音频、视频）】从而形成一个容纳所有形式数据的集中式数据存储，进而为后续的报表、可视化分析、实时分析、以至于机器学习提供数据支撑。

(2) 数据湖就像一个大型容器，与真正的湖泊和河流非常相似。就像在湖中你有多个支流进来一样，数据湖有结构化数据，非结构化数据，机器到机器，实时流动的日志。

(3) 数据湖是一种经济有效的方式来存储组织的所有数据以供以后处理。研究分析师可以专注于在数据中找到意义模式而不是数据本身。

1.3 从数据库、数据仓库到数据湖演变趋势

从 1960 年开始，数据管理经历了数据收集、数据库、数据仓库的阶段，2001 年后随着互联网的迅速发展，大数据时代来临，对数据管理技术提出了全新的要求，未来朝着数据湖的方向演进。

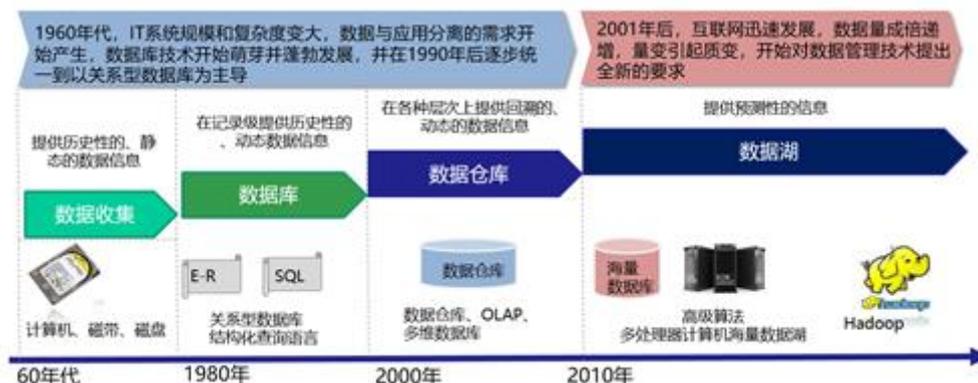


图 3.数据库、数据仓库到数据湖发展历程

数据库的数据有对齐的要求，数据库是面向应用的，每个应用可能需要一个数据库。如果一个公司有几十个应用，就会有几十个数据库。几十个数据库之间怎么去连接分析、统一分析？是没有办法的。

随后就由数据库发展成了一个数据仓库，数据仓库不面向任何应用。但是，它对接到数据库，如果需要每天定时有些 ETL 的批处理的任务，将不同应用和数据汇总起来，按照一些范式模型去做连接分析，得到一定时间段的总体数据视图。这个前提是很多数据库要给数仓供应数据。

而随着数据量的增加及数据类型的变化，很多非结构化的数据，比如视频、音频及文档等占据数据总量的比例越来越多。原来的数据仓库已经很难继续支撑，因此越来越多的企业希望把原始数据以真实的初始状态保留下来。在这种需求的推动下，数据湖的理念便开始成形，其可以把数据保存在原始状态，以便于企业从多个维度进行更多分析。数据可以很轻松进入数据湖，用户也可以延迟数据的采集、数据清洗、规范化的处理，可以把这些延迟到业务需求来了之后再进行处理。传统的数仓，因为模型范式的要求，业务不能随便的变迁，变迁涉及到底层数据的各种变化。相对来说，数据湖就更加的灵活，能更快速的适应上层数据应用的变化。

1.4 数据仓库与数据湖差异

数据湖是按原始数据格式存储，旨在任何数据可以以最原始的形态储存，可是结构化或者非结构化数据，以确保数据在使用时可以不丢失任何细节，所有的实时数据和批量数据，都汇总到数据湖当中，然后从湖中取相关数据用于机器学习或者数据分析。

(1) 相关差异点

·在储存方面上，数据湖中所有数据都保持原始形式，仅在分析时再进行转换。

数据仓库就是数据通常从业务系统中提取。

在将数据加载到数据仓库之前，会对数据进行清理与转换。在数据抓取中，数据湖就是捕获半结构化和非结构化数据。而数据仓库则是捕获结构化数据并将其按模型来组织。

数据湖的目的就是数据湖适合深入分析的非结构化数据。数据科学家可能会用具有预测建模和统计分析等功能的高级分析工具。而数据仓库就是数据仓库非常适用于数据指标、报表、报告等分析用途，因为它具有高度结构化。

数据湖通常在存储数据之后定义架构，较少的初始工作并提供更大的灵活性。而在数据仓库中存储数据之前需定义架构。

特性	数据仓库	数据湖
存储的数据类型	主要处理历史的、结构化的数据，而且这些数据必须与数据仓库事先定义的模型吻合。来自事务系统、运营数据和业务应用程序的关系数据。	能处理所有类型的数据，如结构化数据，非结构化数据，半结构化数据等。数据的类型依赖于数据源系统的原始数据格式。来自IOT设备、网站、移动程序，社交媒体和企业应用程序关系和非关系数据。
数据处理模式	是高度结构化的架构，数据在清洗转换之后才会加载到数据仓库，用户获得的是处理后数据，这叫做写时模式 (Schema-On-Write)。处理结构化数据，将它们或者转化为多维数据，或者转换为报表，以满足后续的高级报表及数据分析需求。	数据直接加载到数据湖中，然后，然后根据分析的需要再处理数据，这叫做读时模式 (Schema-On-Read)。拥有足够强的计算能力用于处理和分析所有类型的数据，分析后的数据会被存储起来供用户使用。
访问方式	数据仓库通常用于存储和维护长期数据，因此数据可以按需访问。	数据湖通常包含更多的相关的信息，这些信息有很高概率会被访问，并且能够为企业挖掘新的运营需求。
性价比	更快的查询结构，存储成本高	更快的查询结构，存储成本低
工作合作方式和服务对象	集中式的，业务人员给需求到数据团队，数据团队根据要求加工、开发成维度表，供业务团队通过BI报表工具查询。 服务对象:业务分析师	是开放、自助式的开放数据给所有人使用，数据团队更多是提供工具、环境供各业务团队使用（不过集中式的维度表建设还是需要的），业务团队进行开发、分析。 服务对象: 数据科学家、数据开发人员和业务分析师。
数据质量	可以作为重要事实依据的高度监管数据	任何可以或者无法监管的数据（例如原始数据）
分析	批处理报告、BI和可视化	机器学习，预测分析，数据发现和分析

图 4.数据仓库和数据湖的差异和联系

(2) 数据湖主要特点

数据湖与数据仓库的理念不同，相对于数据仓库的注重数据管控，数据湖更倾向于于数据服务。

数据湖对数据从业人员的素质要求更高；对数据系统的要求更高，要防止数据湖变数据沼泽，此时就需要借助现代化的数据治理能力。

数据湖与数据仓库不是互斥的。当前条件下，数据湖并不能完全替代数据仓库。尤其是对于已经使用数据仓库的公司，这种情况下数据仓库可以作为数据湖的一个数据来源。

与数据存储在文件和文件夹中的分层数据仓库不同，数据湖具有扁平的架构。数据湖中的每个数据元素都被赋予唯一标识符，并标记有一组元数据信息。

数据湖的三个层次，分为数据库等底层存储、元数据管理、跨不同数据源的 SQL 引擎。

数据湖也是数据仓库发展的高级阶段，对于数据仓库来说，数据湖有很多扩展能力。

数据仓库解决的核心问题，数据湖也解决了一遍，而且涉及面更广。

二、数据湖的架构体系

2.1 数据湖架构体系

数据、算法和算力三大因素正在全力推动数据湖应用快速发展。企业建立统一的数据湖平台，完成数据的采集、存储、处理、治理，提供数据集成共享服务、高性能计算能力和大数据分析算法模型，支撑经营管理数据分析应用的全面开展。为规模化数据应用赋能。

数据湖技术架构涉及了数据接入（转移）、数据存储、数据计算、数据应用、数据治理、元数据、数据质量、数据资源目录、数据安全及数据审计等 10 个方面领域，以下简要作一介绍：



图 5.数据湖包含技术体系

1) 数据接入 (移动)

数据提取允许连接器从不同的数据源获取数据并加载到数据湖中。数据提取支持：所有类型的结构化，半结构化和非结构化数据。批量，实时，一次性负载等多次摄取；在数据接入方面，需提供适配的多源异构数据资源接入方式，为企业数据湖的数据抽取汇聚提供通道。

2) 数据存储

数据存储应是可扩展的，提供经济高效的存储并允许快速访问数据探索。它应该支持各种数据格式。

3) 数据计算

数据湖需要提供多种数据分析引擎，来满足数据计算需求。需要满足批量、实时、流式等特定计算场景。此外，向下还需要提供海量数据的访问能力，可满足高并发读取需求，提高实时分析效率。并需要兼容各种开源的数据格式，直接访问以这些格式存储的数据。

4) 数据治理

数据治理是管理数据湖中使用的数据的可用性，安全性和完整性的过程。数据治理是一项持续的工作，通过阐明战略、建立框架、制定方针以及实现数据共享，为所有其他数据管理职能提供指导和监督。

5) 元数据

元数据管理是数据湖整个数据生命周期中需要做的基础性工作，企业需要对元数据的生命周期进行管理。元数据管理本身并不是目的，它是组织从其数据中获得更多价值的一种手段，要达到数据驱动，组织必须先是由元数据驱动的。

6) 数据资源目录

数据资源目录的初始构建，通常会扫描大量数据以收集元数据。目录的数据范围可能包括全部数据湖中被确定为有价值 and 可共享的数据资产。数据资源目录使用算法和机器学习自动完成查找和扫描数据集、提取元数据以支持数据集发现、暴露数据冲突、推断语义和业务术语、给数据打标签以支持搜索、以及标识隐私、安全性和敏感数据的合规性。

7) 隐私与安全

数据安全是安全政策和安全程序的规划、开发和执行、以提供对数据和信息资产的身份验证、授权、访问和审核。需要在数据湖的每个层中实现安全性。它始于存储，发掘和消耗，基本需求是停止未授权用户的访问。身份验证、审计、授权和数据保护是数据湖安全的一些重要特性。

8) 数据质量

数据质量是数据湖架构的重要组成部分。数据用于确定商业价值，从劣质数据中提取洞察力将导致质量差的洞察力。数据质量重点关注需求、检查、分析和提升的实现能力，对数据从计划、获取、存储、共享、维护、应用、消亡生命周期的每个阶段里可能引发的各类数据质量问题进行识别、度量、监控、预警等一系列活动，并通过改善和提高组织的管理水平使得数据质量获得进一步提高。

9) 数据审计

两个主要的数据审计任务是跟踪对关键数据集的更改：跟踪重要数据集元素的更改；捕获如何/何时/以及更改这些元素的人员。数据审计有助于评估风险和合规性。

10) 数据应用

数据应用是指通过对数据湖的数据进行统一的管理、加工和应用，对内支持业务运营、流程优化、营销推广、风险管理、渠道整合等活动，对外支持数据开放共享、数据服务等活动，从而提升数据在组织运营管理过程中的支撑辅助作用，同时实现数据价值

的变现。在基本的计算能力之上，数据湖需提供批量报表、即席查询、交互式分析、数据仓库、机器学习等上层应用，还需要提供自助式数据探索能力。

2.2 以 AWS 数据湖产品为例，实现数据管理 4 个能力

AWS 数据湖方案主要是基于 AWS 云服务，该方案提出在 AWS 云上部署高可用的数据湖架构，并提供用户友好的数据集搜索和请求控制台。AWS 数据湖方案主要借助了 Amazon S3、AWS Glue、Amazon Athena 等 AWS 服务来提供诸如数据提交、接收处理、数据集管理、数据转换和分析、构建和部署机器学习工具、搜索、发布及可视化等功能。建立以上基础后，再由用户选择其它大数据工具来扩充数据湖。

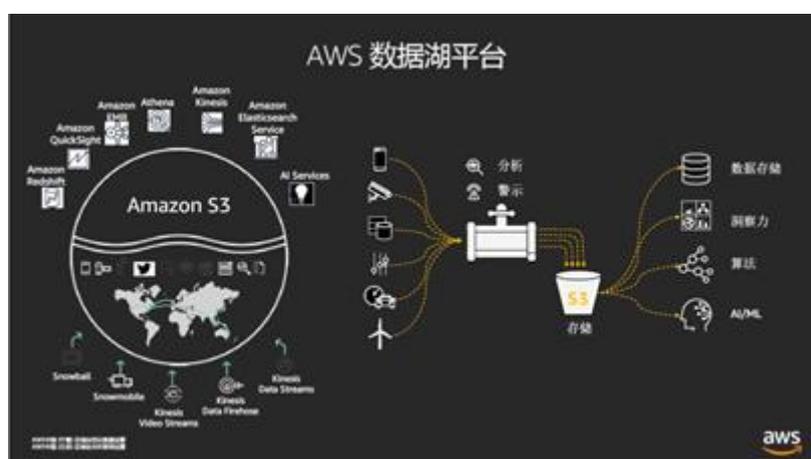


图 6.AWS 数据湖解决方案图

图 6 AWS 数据湖解决方案提供了完整的数据架构支持，为企业构建一站式数据处理体验，目前已在多个行业和客户中使用。例如：数据湖解决方案支撑平安城市“一云一湖一平台”系统架构，为公安客户构建了物理分散（分散在各地市、区县的数据）、逻辑统一的数据治理架构。



图 7.AWS 整个大数据分析服务的全景图

AWS 数据湖的一个典型架构，我们看到数据湖并不是一个产品、也不是一项技术，而是由多个大数据组件、云服务组成的一个解决方案。可以全方位的可以提供最先进的数据湖的大数据分析。

(1) 数据移动组件

数据迁移和移动的工具，有 AWS Database Migration Service (AWS DMS)数据库迁移服务，还有 AWS Snowball (雪球)，像快递一样，可以把数据放在一个专用的硬盘类似的装置里面来快递的服务。还有像混合云里面，AWS Storage Gateway 通过一个数据的门户网关来转换数据，同时推出了 AWS Backup 数据备份服务，这一类的服务是更底层的作为数据移动的服务。在数据移动组件中，还有 Amazon Kinesis 和 Amazon Managed Streaming of Apache Kafka 这些消息队列和流计算工具，其中 Amazon Kinesis 能够轻松收集、处理和分析实时流数据，可以使用 Kinesis Data Firehose 将流式数据持续加载到 Amazon

S3 数据湖中。

近期在中国上线的 AWS Glue 一项全托管的数据提取、转换和加载 (ETL) 服务及元数据目录服务。它能提供完全托管的提取、转换和加载 (ETL)服务，可以用来登记、清理

和丰富数据，并可以在数据存储之间可靠地移动数据，显著降低创建 ETL 任务所花费的费用和时间以及其复杂性。

(2) 数据湖组件

数据湖最主要的元素是三大元素：一个是 Amazon S3/Glacier，一个是 AWS Glue 和 Amazon Athena，一个是 AWS Lake Formation。

- 最核心的组件是 Amazon S3，它可以存储二进制为基础的任何信息，包含结构化和非结构化的数据，例如：企业信息系统 MES、SRM 等系统中的关系型数据，从手机、摄像头拍来的照片、音视频文件，从火力发电机等各种设备产生的数据文件等。借助 Amazon S3，可以通过经济高效的方式构建和扩展任何规模的数据湖。

- 上面提及到的 AWS Glue 服务，还是可以提供数据目录服务的功能。因为数据都存在数据湖里面，在这个过程中，要对这些数据打上标签，把它做分类的工作。Glue 就像爬虫一样对数据湖里的海量数据，进行自动爬取，生成数据目录的功能。而 Amazon Athena 是一种交互式查询服务，让您能够轻松使用标准 SQL 直接分析 Amazon S3 中的数据。

- AWS Lake Formation：把建立数据湖操作的步骤通过工具自动化管理起来，帮助企业在短短的几天的时间完成数据湖的建设工作。

(3) 数据分析组件

Amazon Redshift 是数据仓库，Amazon EMR 是大数据分析，AWS Glue 在里面仍起关键作用，来实现无服务器的数据分析，然后是 Amazon Athena (雅典娜) 是做交互式的分析，Amazon Elasticsearch 是做一些运维分析，还有 Amazon Kinesis 做实时的数据分析。

- Amazon Redshift 是世界上速度最快的云数据仓库，并且速度每年都在提高。对于性能密集型工作负载，您可以使用新的 RA3 实例将任何云数据仓库的性

能提高多达 3 倍。Redshift Spectrum 直接在 Amazon S3 数据湖中查询数据的功能，客户只需数小时而不是数天或数周，就能轻松整合新的数据源。

·Amazon Athena 是一种交互式查询服务，让您能够轻松使用标准 SQL 分析 Amazon S3 中的数据。只需指向存储在 Amazon S3 中的数据，定义架构并使用标准 SQL 开始查询。就可在数秒内获取最多的结果。使用 Athena，无需执行复杂的 ETL 作业来为数据分析做准备。

(4) 机器学习组件

Amazon SageMaker 是一个人工智能的服务，把这些大数据用来做机器学习、人工智能的数据分析，做更多的自动的预测性的分析。

Amazon SageMaker 也是一项完全托管的服务，可以帮助开发人员和数据科学家快速构建、训练和部署机器学习(ML)模型。SageMaker 完全消除了机器学习过程中每个步骤的繁重工作，让开发高质量模型变得更加轻松。

2.3 数据湖数据管理 4 个能力

AWS 数据湖在数据处理、数据分析、数据服务和安全管控四个方面能力。

(1) 数据处理层面

在 AWS 上轻松运行 Spark, Hadoop, Hive, Presto, Hbase 等大数据分析，更多要使用实时的数据，原来更多的是批量的历史数据，处理实时数据服务叫 Amazon Kinesis，还有四个不同的类型，有的是直接处理视频的数据流，有的是可以把数据直接导到关键的服务，每个各自都有不同的用法。端到端实时建模、跨引擎建模、流式建模等能力优化存储效率，提升存储能力、高效的内存计算能力和高并发数据处理能力。

(2) 数据分析层面

AWS Glue 来实现无服务器的数据分析， Amazon Athena 是做交互式的分析， Amazon Elasticsearch 是做一些运维分析， Amazon Kinesis 做实时的数据分析。实

现六个转变：无服务器分析，提供按需数据湖分析转变、从统计分析向预测分析转变、从被动分析向主动分析转变、从非实时向实时分析转变、从结构化数据向多元化转变。

(3) 提供多种数据服务

提供统一的、标准的数据服务，数据资产可知、可查、可用，要有资产清单、数据资产共享需要授权和流程的管控。

- ETL 和数据目录服务；

- 人工智能服务：帮助开发人员将预先构建的人工智能功能插入到他们的应用程序中；

- 机器学习平台服务：帮助所有开发人员轻松入手并深入了解机器学习。

(4) 数据安全及管控层面

Amazon S3、Amazon DynamoDB、Amazon Redshift 具备很好的数据安全机制，数据的传输和存储都是加密的，加密密钥只有客户自己掌握，防止数据泄露带来的风险，保障数据共享的安全。另外，还有 Amazon VPC 安全策略、AWS IAM、AWS KMS 等安全组件为 AWS 数据湖保驾护航，为企业数据的存储、处理、使用提供一个安全、合规的数据环境，平台管控要可视化，提高运维效率，实现统一的数据流监控，降低运维成本。

三、 如何通过数据治理实现数据湖商业价值

数据湖对一个企业的数字化转型和可持续发展起着至关重要的作用。构建开放、灵活、可扩展的企业级统一数据管理和分析平台，将企业内、外部数据按需关联，打破了数据的系统界限。

- 1) 利用数据湖智能分析、数据可视化等技术，实现了数据共享、日常报表自动生成、快速和智能分析，满足企业各级数据分析应用需求。

- 2) 深度挖掘数据价值，助力企业数字化转型落地。实现了数据的目录、模型、标准、认责、安全、可视化、共享等管理，实现数据集中存储、处理、分类与管理，实现报

表生成自动化、数据分析敏捷化、数据挖掘可视化，实现数据质量评估、落地管理流程。

3.1 数据湖遇到挑战

数据湖本身是一个中心化的存储，能够存储任意规模的结构化与非结构化数据。数据湖的优势就是数据可以先作为资产存放起来，问题就在于如何把这些数据在业务中利用起来。当部署了数据湖之后，数据治理问题将会接踵而至，比如从数据湖到数据湖，如何将数据进行分流、湖的数据如何进行整理等。

数据仓库里的数据是经过整理、清晰易懂的。而数据湖的概念是不经处理直接进行堆砌，那么数据湖就有可能变成“数据沼泽”，筛选难度会变大。由于定义不正确、信息不完整、数据陈旧或无法找到所需信息，它需要更多的元数据来理解存储在数据湖中的数据资产，包括数据内容、数据资产图谱、数据敏感性、用户喜好、数据质量、上下文（缺乏上下文将无法用于分析）和数据价值等业务层面的理解。另外这些系统和应用是技术人员开发的，由于技术人员和业务人员的思维和“语言”存在差异，这使得业务用户获取数据变得更加复杂和困难。

3.2 避免数据沼泽

如何让数据湖的水保持清亮不会成为数据沼泽？“数据湖的数据不被有效使用就会成为大垃圾场。”中国有句谚语：“流水不腐，户枢不蠹”。数据只有流动起来，才可以不成为数据沼泽，湖泊只是暂存数据河流的基地。数据流动就意味着所有的数据产生，最终要有它的耕种者和使用者。要让数据有效流动起来，就要建立有效的“数据河”（Data River）。

业界在数据湖的尝试上一般都会忽视数据治理的重要性，这是很危险的，由它导致的数据沼泽也是企业对数据湖持续观望的原因之一。

3.3 数据智能化治理是数据湖实现价值必有之路

对数据治理的需求实际更强了。因为与“预建模”方式的数仓不同，湖中的数据更加分散、无序、不规则化等，需要通过治理工作达到数据“可用”状态，否则数据湖很可能会“腐化”成数据沼泽，浪费大量的 IT 资源。平台化的数据湖架构能否驱动企业业务发展，数据治理至关重要，没有数据湖治理，企业可能失去有意义的商业智能。这也是对数据湖建设的最大挑战之一。

数据湖以数据治理为基础、建立一套自助服务为抓手的工具链来赋能业务发展。数据湖能给企业带来多种能力，例如，能实现数据的集中式管理，在此之上，企业能挖掘出很多之前所不具备的能力。另外，数据湖结合先进的数据科学与机器学习技术，能帮助企业构建更多优化后的运营模型，也能为企业提供其他能力，如预测分析、推荐模型等，这些模型能刺激企业能力的后续增长。

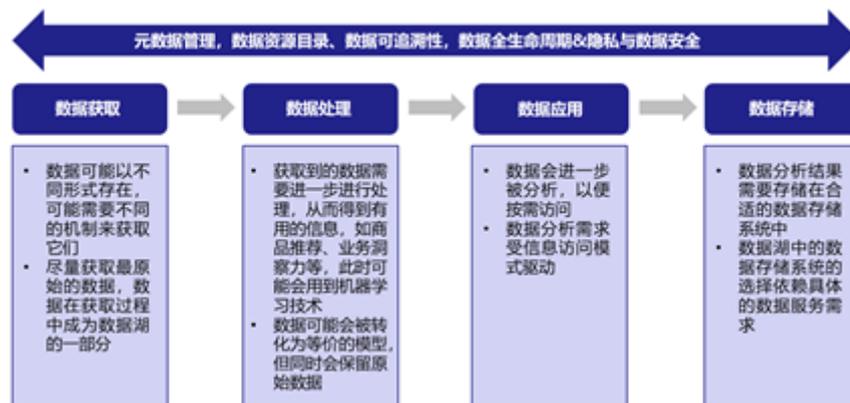


图 8.数据湖中数据全生命周期管理

当数据从采集点流入数据湖时，它的元数据被捕获，并根据其生命周期中的数据敏感度从数据可追溯性、数据全生命周期和数据安全等方面进行管理。在数据大爆发的背景下，数据治理对数据湖起到关键作用，因为数据治理涉及组织中跨功能和跨业务的所有决策机制。数据智能在提供数据支持和数据治理应用方面至关重要，因为它为企业提供了在最佳时间内将正确的数据交付给正确的对象所需的知识。数据智能也在帮助专业人士在工作中变得更高效、更有效，在可靠数据的支持下做出更好的数据驱动决策。

3.4 构建数据湖的数据治理体系相关思考

笔者认为，数据湖的数据治理体系包括元数据管控、数据资源目录、主数据管控、数据服务、数据全生命周期管理、数据质量提升及隐私与安全管理等内容。而这只是数据湖管理难题的一部分。考虑全面的数据湖治理，包括是谁引入的数据、谁负责数据，以及数据的定义，以确保数据的妥善标记和使用，实现对企业数据资源内容层面的优化改造和有效管控。

(1) 元数据管控

传统的数据仓库将数据存储于关系表中，而数据湖则使用平面结构。每个数据元素被分配唯一标识符，并用一组元数据标签进行标记。这就是说，数据湖没有数据仓库那么结构化。

设计元数据标准及采集方案、元数据应用、管理流程等，形成企业级数据资源目录与全链式数据流通追踪，实现对企业数据资源的清晰掌握和数据流通全流程的监控，满足分布式部署模式下数据资源完整性管理及应用的需求。

数据湖解决方案为企业中海量的数据集提供了一套集中的元数据管理系统，提供全局的数据资源目录、完整的数据元数据描述、数据血缘关系，方便员工快速查找了解数据，更好的支撑数据分析。

(2) 数据资源目录

数据资源目录包含业务术语表关联、标签管理、数据分类、数据来源和全文检索。通过最大限度的自动化和有限的人工操作，可以从构建的数据资产目录中获得更多价值。例如利用机器学习可以实现数据自动分类和打标签。再如，有监督学习技术是基于已经打上标签的样本数据上训练一个模型，然后将该模型应用于所有未打标的数据，在这些数据中，实例根据预测中的信任度进行排序。最自信的预测然后被添加到标记的例子中。这个过程不断重复，直到所有未标记的例子都被标记。

(3) 主数据管控

面向数据湖内全量数据，基于数据关系，实现自动化的主数据识别映射、主数据一致性维护、主数据关系发布等功能，搭建企业核心业务对象数据的管理体系，支撑跨业务的数据联动以及基于数据驱动的业务协同。

(4) 数据质量提升

针对企业缺乏对全部数据资源进行系统质量控制的现状，设计企业级数据质量规则定义、控制管理流程和手段，提高并确保数据质量，为业务应用提供规范、准确的数据支撑。有效的数据湖部署需要数据质量分析师、工程师与数据治理团队、数据管理员密切合作，以部署数据质量策略、分析数据并采取必要的措施来提高其质量。

(5) 数据全生命周期管理

数据的生命周期，包括数据的起源以及数据是如何随时间移动的。它描述了数据在各种处理过程中发生了哪些变化，有助于提供数据分析流水线的可见性，并简化了错误溯源。通过对元数据的关系解析和血缘分析，构建全维关系图谱，实现关系融合。通过对数据的血缘分析、数据标签等方法，实现数据多版本共存条件下的统一身份和可控的数据归一化，最终实现的数据全生命周期管理和追踪。

(6) 数据服务

1) 主题数据服务

面向企业分析应用，提供按业务主题的数据组织能力，支撑企业生产管理与经营决策的业务主题构建和分析需求。

2) AI 数据服务

为 AI 分析引擎创建探索数据，构建基础标签体系，提供快速、全量的数据支撑。

3) 微服务数据服务

按照云端 SaaS 应用的开发部署模式和弹性部署需求，构建微服务数据组织能力，发布数据服务 API，实现应用与数据的松耦合。

(7) 数据质量提升

有效的数据治理使企业能够提高数据湖中的数据质量，并利用数据进行业务决策，从而可以改善业务规划和财务绩效，因此定义数据源以及管理和使用数据至关重要。企业还可以考虑在消费方而不是采购方应用数据质量检查。因为，单个数据质量体系结构可能不适用于所有类型的数据。必须注意的是，如果数据被“清理”，用于分析的结果可能会产生影响。修复数据集中值的字段级数据质量规则可以影响预测模型结果，因为这些修复可以影响异常值。

(8) 隐私与安全

数据安全标准和策略未被正确纳入治理流程中，可能会导致无法访问受隐私法规和其他类型的敏感数据保护的个人信息。健康数据湖的关键组成部分是隐私和安全性，包括基于角色的访问控制、身份验证、授权以及静态和动态数据加密等。从纯数据湖和数据管理的角度来看，最重要的往往是数据混淆，包括标记化和数据屏蔽。应该使用这两个概念来帮助数据遵守最小特权的安全概念。限制数据访问也对许多希望遵守法规的企业具有意义。尽管数据湖旨在成为相当开放的数据源，但仍需要安全性和访问控制措施，数据治理和数据安全团队应携手完成数据湖设计和加载过程，以及持续的数据治理工作。

四、 Amazon Athena 和 AWS Glue 中国区域实践案例

AWS Glue 现已在由光环新网运营的 AWS 中国（北京）区域和由西云数据运营的 AWS 中国（宁夏）区域正式上线。AWS Glue 是一项全托管的数据提取、转换和加载 (ETL) 服务及元数据目录服务。它让客户更容易准备数据，加载数据到数据库、数据仓库和数据湖，用于数据分析。使用 AWS Glue，在几分钟之内便可以准备好数据用于分析。由于 AWS Glue 是无服务器服务，客户在执行 ETL 任务时，只需要为他们所消耗的计算资源付费。

同时在中国上线的还有 Amazon Athena，它是一种交互式查询服务，让客户可以使用标准 SQL 语言、轻松分析 Amazon S3 中的数据。由于 Athena 是一种无服务器服

务，因此客户不需要管理基础设施，而且只为他们运行的查询付费。Athena 可以自动扩展，并行执行查询，所以即便是大型数据集和复杂的查询，也能很快获得查询结果。

4.1 ETL 服务为数据分析准备工作的自动化，大幅缩短数据准备时间

全新的 ETL 服务实现了数据分析准备工作的自动化，让客户从准备数据到开始分析的时间由几个月缩短到几分钟。客户在使用数据湖架构实现数据分析解决方案时，通常有 75%的时间花在数据集成任务上，需要从各种数据源提取数据，对其进行规范化，并将其加载到数据存储中。AWS Glue 消除了 ETL 作业基础设施方面的所有重复劳动，让 Amazon S3 数据湖中的数据集可以被发现、可用于查询和分析，极大地缩短分析项目中做 ETL 和数据编目阶段的时间，让 ETL 变得很容易。通过简化创建 ETL 作业的过程，AWS Glue 让客户可以构建可伸缩、可靠的数据准备平台。这些平台可以跨越数千个 ETL 作业，具有内置的依赖性解析、调度、资源管理和监控功能。

4.2 数据资源目录为数据湖提供智能化数据管理能力

AWS Glue 数据资源目录功能可以通过一个爬虫直接获取在 Amazon S3 上的数据目录，用于查询。在从客户选择的数据源把数据爬取出来之后，会自动识别数据格式和模式 (schema)，构建统一的数据目录，并为客户提供所选数据的中央视图。这使得客户很容易跨越各种数据存储，检索和管理所有数据，而不必手动搬运它们。当客户从数据目录中标识出数据源 (例如一个数据库表) 和数据目标 (例如一个数据仓库) 时，AWS Glue 将匹配相应的模式，生成可定制、可重用、可移植、可共享的数据转换代码。AWS Glue 的数据目录功能让客户可以轻松使用 Amazon Elastic MapReduce (Amazon EMR) 来直接处理和查询 Amazon S3 上的数据，提高了企业的开发效率。

4.3 交互式查询服务为数据湖提供高效、便捷服务能力

通过 Amazon Redshift，客户可以对大规模的结构化数据执行复杂的查询，并获得超高速的性能。对于非结构化数据，Amazon EMR 使用流行的分布式框架，例如

Apache Spark、Presto、Hive 和 Pig，横跨多个可动态伸缩的集群，处理和分析大量数据，快速又经济。使用 Athena 分析 Amazon S3 中的数据就像编写 SQL 查询一样简单。Athena 使用完整支持标准 SQL 的 Presto，可以处理各种标准数据格式，包括 CSV、JSON、ORC 和 Parquet。因为 Athena 使用多个可用区的计算资源执行查询，而且使用 Amazon S3 作为底层数据存储，所以它具有高可用性和持久性，数据冗余存储在多处基础设施中，并且是每处基础设施上的多个设备上。

五、数据湖的未来展望

现阶段数据湖更多是作为数据仓库的补充，它的用户一般只限于专业数据科学家或分析师。数据湖概念和技术还在不断演化，不同的解决方案供应商也在添加新的特性和功能，包括架构标准化和互操作性、数据治理要求、数据安全性等。

数据湖作为一种云服务随时按需满足对不同数据的分析、处理和存储需求，数据湖的扩展性，可以为用户提供更多的实时分析，基于企业大数据的数据湖正在向支持更多类型的实时智能化服务发展，将会为企业现有的数据驱动型决策制定模式带来极大改变。

六、结束语

在数字经济时代里，从数据仓库到数据湖，不仅仅是数据存储架构的变革，更是大数据思维方式的升级。用好数据是企业数字化转型的关键、数据湖是数据分析智能商务的新趋势。数据湖能给企业带来多种能力，数据湖结合先进的数据科学与机器学习技术，能帮助企业构建更多优化后的运营模型，也能为企业提供其他能力。数据湖将以数据治理为基础、依托一套自助服务为抓手的工具链来赋能业务发展。